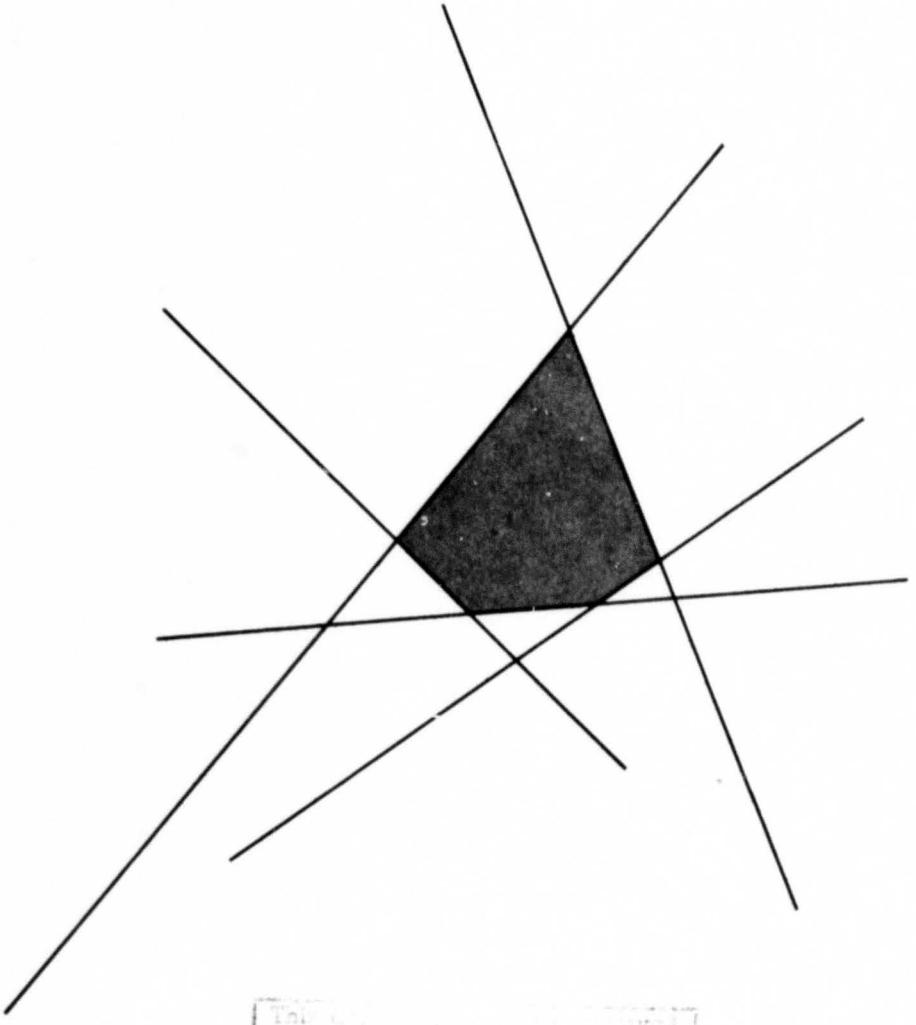
ORC 69-8
APRIL 1969

NONZERO-SUM STOCHASTIC GAMES

by

PHILIP D. ROGERS



OPERATIONS RESEARCH CENTER

UNIVERSITY OF CALIFORNIA · BERKELEY

45

NONZERO-SUM STOCHASTIC GAMES

by

Philip D. Rogers
Operations Research Center
University of California, Berkeley

APRIL 1969 ORC 69-8

This research has been supported by the Office of Naval Research under Contract Nonr-222(83) with the University of California. Reproduction in whole or in part is permitted for any purpose of the United States Government.

ACKNOWLEDGEMENT

I thank Professors Ronald W. Shephard, John C. Harsanyi, and Richard E. Barlow for their help in preparing this paper. Steve Jacobsen, Bill Mitchell, and K. G. Murty also have my thanks for the time they spent in useful discussions about the problems herein.

I appreciate the efforts of Linda Betters, Noreen Comotto, and Cathy Korte who did an excellent typing job. Finally, I thank the National Science Foundation for their support during my graduate studies.

ABSTRACT

This paper extends the basic work that has been done on zero-sum stochastic games to those that are nonzero-sum. Appropriately defined equilibrium points are shown to exist for both the case where the players seek to maximize the total value of their discounted period rewards and the case where they wish to maximize their average reward per period. For the latter case, conditions required on the structure of the Markov chains are less stringent than those imposed in previous work on zero-sum stochastic games, extensions to n-person games and underlying semi-Markov processes are discussed, and finding an equilibrium point is shown to be equivalent to solving a certain nonlinear programming problem.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

A stochastic game combines a finite state, discrete time sequential decision process with two person game theory in the following way: at time n, two players are jointly in some state i, $i = 1, \ldots, N$, in which they play a $K_i \times L_i$ bimatrix game $[A^i, B^i]$. If the players choose row k and column ℓ respectively, then $a^i_{k\ell}$ is the reward to player I and $b^i_{k\ell}$ the reward to player II. The players' choices also determine $p^{k\ell}_{ij}$, the probability that the players move from state i to state j at time $n+1$, $j = 1, \ldots, N$.

A stationary strategy for player I in state i is a probability vector $x_i = (x_{i1}, x_{i2}, \ldots, x_{iK_i})$ where x_{ik} is the probability that player I chooses the kth row, and player I uses x_i whenever in state i. Similarly, a stationary strategy for player II in state i is a probability vector $y_i = (y_{i1}, y_{i2}, \ldots, y_{iL_i})$ where $y_{i\ell}$ is the probability that player II chooses the ℓth column, and player II uses y_i whenever in state i. If the players have chosen strategies x_i and y_i, then player I's expected reward for period n is

$$\sum_{k=1}^{K_i} \sum_{\ell=1}^{L_i} a^i_{k\ell} x_{ik} y_{i\ell}$$

and player II's expected reward is

$$\sum_{k=1}^{K_i} \sum_{\ell=1}^{L_i} b^i_{k\ell} x_{ik} y_{i\ell} .$$

At time $n+1$, the players will be in state j with probability

$$p_{ij} = \sum_{k=1}^{K_i} \sum_{\ell=1}^{L_i} p_{ij}^{k\ell} x_{ik} y_{i\ell}$$

where the bimatrix game $[A^j, B^j]$ will be played, stationary strategies x_j , y_j employed, and a transition to a new state made. The game continues in this manner over an infinite horizon, the movement of the players being governed by the Markov chain (p_{ij}) .

There are several possibilities for the objectives of the two players. We will first study the case in which the players seek stationary strategies $x = (x_1, \ldots, x_N)$ and $y = (y_1, \ldots, y_N)$ respectively which will uniformly maximize, for all initial states, the discounted value of their total expected rewards. Then we will examine the case in which the players desire to maximize their expected reward per period, and seek stationary strategies to do so. These will be referred to as the discounted case and average rate of return case, respectively. It is clear that what is good for one player may be bad for the other, so it will generally be impossible for both players to simultaneously achieve these objectives (in the zero-sum game, this is always the case since with $A^i = -B^i$, the players have directly opposing interests). Hence, we turn to the concepts of a "value" for a zero-sum stochastic game and an "equilibrium point" for a nonzero-sum stochastic game, discussed in Chapter 2.

The literature on stochastic games is not extensive. The first article appeared in 1953, when L. S. Shapley [15] first described the game. Shapley proved the existence of an appropriately defined value for a zero-sum game with total discounted rewards as the payoffs. He showed that an optimal strategy that achieves the value can be taken to be stationary, i.e., the players can use the same strategies every time they are in state i independent of the time period in which they arrive in state i , and he

provided an algorithm for the determination of optimal strategies and the value.

The average rate of return zero-sum game was treated by D. Gillette [4] in 1957. Whereas the structure of the Markov chain governing the transitions of the players can be arbitrary in the discounted zero-sum game, Gillette showed that this is not the case when average rate of return is the objective if we hope to have stationary strategies yield a value for the game. He accomplished this by proving that if all possible underlying Markov chains are irreducible, then a value exists and can be achieved by stationary strategies, and he gave an example of a game having a reducible chain for which a value could not be attained by stationary strategies.

Gillette's results were rederived from a linear programming approach by Hoffman and Karp [6]. Their results required the retainment of the irreducibility assumption. In addition, they presented an algorithm which converges to stationary strategies yielding the value of the game.

The results that follow generalize those above to nonzero-sum stochastic games and provide a relaxation of the irreducibility assumption in the average rate of return case. Following the work of Nash [13] on nonzero-sum games, the existence of appropriately defined equilibrium points for nonzero-sum stochastic games is proven for both the discounted and average rate of return games. In the latter case, the irreducibility assumption is weakened to allow for some transient states as long as every possible underlying chain has a single ergodic subclass of states. For the average rate of return case, an equilibrium point is shown to be equivalent to solving a nonlinear programming problem and extensions to n-person games and underlying semi-Markov processes discussed. As a biproduct of these efforts in the discounted case and average rate of return case with irreducible chains, we get a

characterization of the set of stationary optimal policies for a sequential decision process, the process that results from letting one of the players be a "dummy" with only one possible action available in each state.

CHAPTER 2

DISCOUNTED CASE

2.1 Introduction

Since a nonzero-sum stochastic game can be viewed as the marriage of a nonzero-sum game and a discrete dynamic programming problem (sequential decision process), it comes as little surprise that the major results for such games depend heavily on the results and structure of both these subjects. Underlying this relationship is the fact that the major element in proving the existence of equilibrium points for nonzero-sum games is the character of the set of optimal strategies for one player when opposing a given stationary strategy of the other.† But in a stochastic game, when a player's opponent fixes his strategy, the player is faced with precisely a sequential decision process.

In the following two sections, reviews of Nash's work on nonzero-sum games [13] and discrete dynamic programming will be presented and notation set up. Then, in 2.4, the results from these areas will be put together to establish the existence of an equilibrium point for a nonzero-sum stochastic game with expected discounted totals the objective.

2.2 Bimatrix Games

Consider a two-person nonzero-sum bimatrix game $[A,B]$, where A and B are $K \times L$ matrices, $K, L < \infty$, $(A)_{i,j} = a_{ij}$, $(B)_{i,j} = b_{ij}$. Player I (the "row player") has K pure strategies $e_1, e_2, \ldots, e_K$ where e_k is the kth unit vector and player I's use of e_k represents his choice of the kth row of the matrices A and B with probability 1. Similarly, player II

†The games considered are two person unless otherwise indicated.

(the "column player") has L pure strategies $e_1, e_2, \ldots, e_L$ where player II's use of e_ℓ represents his choice of the ℓth column of A and B with probability 1. Corresponding to each pair of pure strategies (e_k, e_ℓ), one strategy being taken for each player, are the rewards $a_{k\ell}$ and $b_{k\ell}$ to players I and II respectively. Mixed strategies $x = (x_1, x_2, \ldots, x_K)$ and $y = (y_1, y_2, \ldots, y_L)$ represent probability distributions over the choices of pure strategies for the players, and when employed, result in expected reward $xAy = \sum_{k=1}^{K} \sum_{\ell=1}^{L} a_{k\ell} x_k y_\ell$ for player I and expected reward $xBy = \sum_{k=1}^{K} \sum_{\ell=1}^{L} b_{k\ell} x_k y_\ell$ for player II.

A pair of strategies (x^o, y^o) is said to be an "equilibrium point" if x^o maximizes xAy^o and y^o maximizes x^oBy. The appealing aspect of an equilibrium point is the stability of such a point in the sense that each player can do no better than to use his equilibrium strategy when opposing the equilibrium strategy of the other. (For a discussion of equilibrium points, their properties and drawbacks, see Luce and Raiffa [10].)

Nash set up the problem of establishing the existence of an equilibrium point for the above game by forming a closely associated correspondence whose fixed points are precisely the equilibrium points for the game. Let

$$X = \left\{ x \mid x \in E^K , \sum_{k=1}^{K} x_k = 1 , x_k \geq 0 \right\} \text{ be player I's strategy space}$$

$$Y = \left\{ y \mid y \in E^L , \sum_{\ell=1}^{L} y_\ell = 1 , y_\ell \geq 0 \right\} \text{ be player II's strategy space}$$

$$\phi_1(\bar{y}) = \left\{ \bar{x} \mid \max_{x \in X} xA\bar{y} = \bar{x}A\bar{y} \right\}$$

$$\phi_2(\bar{x}) = \left\{ \bar{y} \mid \max_{y \in Y} \bar{x}By = \bar{x}B\bar{y} \right\}.$$

Note:

$$\phi_1 \times \phi_2 : Y \times X \to 2^{Y \times X} .$$

Now $(x^o,y^o) \in \phi_1(y^o) \times \phi_2(x^o) \iff x^o \in \phi_1(y^o)$ and $y^o \in \phi_2(x^o)$. Hence (x^o,y^o) is an equilibrium point of the game $[A,B]$ if and only if (x^o,y^o) is a fixed point of $\phi_1 \times \phi_2$. Having established the correspondence between equilibrium points of the game and fixed points of the correspondence $\phi_1 \times \phi_2$, it only remains to prove the existence of a fixed point for $\phi_1 \times \phi_2$. Since X and Y are nonempty, compact and convex, this can be accomplished by Kakutani's fixed point theorem [9] which requires that $\phi_1 \times \phi_2$ have a closed graph[†] and that $\phi_1(y) \times \phi_2(x)$ be convex and nonempty for all $(y,x) \in Y \times X$, all of which hold.

2.3 Sequential Decision Processes

Consider the classical sequential decision process with an infinite planning horizon and discount factor β, $0 \leq \beta < 1$.[††] At the beginning of period n $(n = 1,2, \ldots)$, a player (decision maker) finds himself in one of a finite number of states $\{1,2, \ldots, N\}$, say i, and is faced with choosing one of a finite number of actions $\{1,2, \ldots, K_i\}$. As a consequence of choosing action k, the player experiences an immediate expected reward, r_{ik}, and a transition to a new state j, the latter occurring with probability p_{ij}^k, $\sum_{j=1}^{N} p_{ij}^k = 1$. Note that both his reward and the probabilities governing his movement depend on the state he's in (i) and the action he chooses (k).

A randomized stationary strategy $x_i = \left(x_{i1}, x_{i2}, \ldots, x_{iK_i}\right)$, $i = 1,2, \ldots, N$, is simply a set of N probability vecotrs where, every time the player is in state i, x_{ik} is the probability that he chooses action k. It follows that the use of x_i in state i will result in an

---

[†] A correspondence $\phi : U \to V$ is said to have a closed graph if for every sequence $u^q \to u^o$ and $v^q \to v^o$ with $v^q \in \phi(u^q)$ $\forall$ q, we have $v^o \in \phi(u^o)$.

[††] β^n is the present value of a unit reward earned n periods in the future.

immediate expected reward

$$r_i(x) = \sum_{k=1}^{K_i} r_{ik} x_{ik}$$

and a transition to a new state j with probability

$$p_{ij}(x) = \sum_{k=1}^{K_i} p_{ij}^k x_{ik} .$$

Hereafter, the word "strategy" will mean "stationary strategy."

$V(x)$ is defined to be a column vector whose ith component, $V_i(x)$, is the expected total reward over all future time, discounted to the beginning of a period when the player is in state i, and strategy x is employed. It is clear that $V(x)$ satisfies

$$V(x) = r(x) + \beta P(x)V(x) \tag{1}$$

where $r(x)$ is a column vector whose ith component, $r_i(x)$, is the immediate expected reward in state i and $P(x)$ is the Markov chain, whose ith row, $P_i(x)$, governs transitions from state i, when strategy x is employed. From (1) we get

$$V(x) = [I - \beta P(x)]^{-1} r(x) , \tag{2}$$

the inverse of $[I - \beta P(x)]$ guaranteed since $0 \leq \beta < 1$.

The strategy x^* is said to be optimal if it maximizes $V(x)$, i.e., if for any strategy x, $V_i(x^*) \geq V_i(x)$, $i = 1, \ldots, N$. It is well known that in the class of randomized strategies, such an optimal strategy exists. (See Hadley [5].)

2.4 Existence of Equilibrium Points in Discounted Case

Using the method of 2.2 and the structure of the sequential decision process of 2.3, we wish to prove that an equilibrium point in stationary strategies exists for the discounted case of a nonzero-sum stochastic game. In order to establish this result, it will be useful to show explicitly that when player II uses some fixed stationary strategy $\bar{y}$, player I is faced with exactly the sequential decision process discussed in 2.3. To see this, suppose player II employs $\bar{y}$. Then if player I chooses action k when in state i, his immediate reward will be

$$a_{ik}(\bar{y}) = \sum_{\ell=1}^{L_i} a_{k\ell}^{i} \bar{y}_{i\ell}$$

and the players will move to state j with probability

$$p_{ij}^{k}(\bar{y}) = \sum_{\ell=1}^{L_i} p_{ij}^{k\ell} \bar{y}_{i\ell} ,$$

exactly the situation of a sequential decision process. Player I's total discounted reward vector now depends upon the strategy of his opponent, $\bar{y}$, as well as his own and he will seek to maximize $V(x,\bar{y}) = [I - \beta P(x,\bar{y})]^{-1} a(x,\bar{y})$. Similar comments apply to player II and his attempt to maximize his total value vector $W(\bar{x},y) = [I - \beta P(\bar{x},y)]^{-1} b(\bar{x},y)$, when player I uses strategy $\bar{x}$.

Let

$$X = \left\{ x \mid x = (x_1, x_2, \ldots, x_N) , x_i \in E^{K_i} , \sum_{k=1}^{K_i} x_{ik} = 1 , x_{ik} \geq 0 \right\}$$

be player I's strategy space,

$$Y = \left\{ y \mid y = (y_1, y_2, \ldots, y_N) \ , \ y_i \in E^{L_i} \ , \ \sum_{\ell=1}^{L_i} y_{i\ell} = 1 \ , \ y_{i\ell} \geq 0 \right\}$$

be player II's strategy space,

$$\theta_1(\bar{y}) = \left\{ \bar{x} \mid \max_{x \in X} V(x,\bar{y}) = V(\bar{x},\bar{y}) \right\}$$

$$\theta_2(\bar{x}) = \left\{ \bar{y} \mid \max_{y \in Y} W(\bar{x},y) = W(\bar{x},\bar{y}) \right\}$$

$$\theta_1 \times \theta_2 : Y \times X \to 2^{Y \times X} \ .$$

Definition:

The pair of strategies (x^o, y^o) is said to be an equilibrium point if $x^o \in \theta_1(y^o)$ and $y^o \in \theta_2(x^o)$.

For any pair of strategies, $(\bar{x},\bar{y})$, both $\theta_1(\bar{y})$ and $\theta_2(\bar{x})$ are nonempty. So following 2.2, if it can be shown that $\theta_1 \times \theta_2$ is convex and has a closed graph, then Kakutani's fixed point theorem can be applied and the existence of an equilibrium point established.

Lemma 1:

$\theta_1 : Y \to 2^X$ and $\theta_2 : X \to 2^Y$ have closed graphs.

Proof:

A sufficient condition for θ_1 to have a closed graph is the continuity of $V(x,y)$. This is assured since the inverse of $[I - \beta P(x,y)]$ always exists and its elements are ratios of polynomials involving the x_{ik} and $y_{i\ell}$ while the elements of $a(x,y)$ are just bilinear terms in the x_{ik} and $y_{i\ell}$. An identical argument on $W(x,y)$ and $b(x,y)$ yields the closed graph nature of θ_2 .

Lemma 2:

$\theta_1(\bar{y})$ can be characterized as a closed convex polyhedron whose extreme points constitute the set of pure strategies that are optimal responses to $\bar{y}$.

Proof:

Since player I is faced with a sequential decision process when player II fixes his strategy at $\bar{y}$, it is necessary and sufficient to show that, for a sequential decision process: (a) any convex combination of optimal pure strategies is an optimal randomized strategy and (b) if some randomized strategy is optimal, then it is a convex combination of optimal pure strategies.

(a) Suppose x^1 and x^2 are pure strategies that are optimal. We want to show that $x_\lambda = \lambda x^1 + (1 - \lambda)x^2$, $0 \leq \lambda \leq 1$, is also optimal. We know:

$$V(x^1) = r(x^1) + \beta P(x^i)V(x^1) \tag{1}$$

$$V(x^2) = r(x^2) + \beta P(x^2)V(x^2) \tag{2}$$

$$V(x^1) = V(x^2) \ . \tag{3}$$

Now consider the strategy which consists of using x_λ for the first period and x^1 thereafter. The total reward for such a strategy is:

$$r(x_\lambda) + \beta P(x_\lambda)V(x^1) = r[\lambda x^1 + (1 - \lambda)x^2] + \beta P[\lambda x^1 + (1 - \lambda)x^2]V(x^1)$$

$$= \lambda[r(x^1) + \beta P(x^1)V(x^1)] + (1 - \lambda)[r(x^2) + \beta P(x^2)V(x^1)]$$

$$= \lambda V(x^1) + (1 - \lambda)V(x^2) = V(x^1)$$

the last two equalities following from (1), (2) and (3). Hence, using x_λ the first period and x^1 thereafter achieves the same total value vector as the optimal strategy x^1. Hadley [5] has shown that this is sufficient to imply that the stationary strategy x_λ is also optimal.

(b) First express an optimal randomized strategy, x, as a convex combination of pure strategies. This can always be done as follows: Let $e_{k_1 k_2 \cdots k_N} = \left(e_{k_1}, e_{k_2}, \ldots, e_{k_N}\right)$ be the pure strategy which chooses alternative k_i in state i. Then

$$(4) \qquad x = \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} \cdots \sum_{k_N=1}^{K_N} x_{1k_1} x_{2k_2} \cdots x_{Nk_N} e_{k_1 k_2 \cdots k_N} .$$

To see why this representation is correct, consider the first K_1 coordinates of the right-hand side and write as

$$\sum_{k_1=1}^{K_1} x_{1k_1} \left[\sum_{k_2=1}^{K_2} \cdots \sum_{k_N=1}^{K_N} x_{2k_2} \cdots x_{Nk_N} \cdot e_{k_1} \right] .$$

Since

$$\sum_{k_2=1}^{K_2} \cdots \sum_{k_N=1}^{K_N} x_{2k_2} \cdots x_{Nk_N} = 1 ,$$

this sum becomes

$$\sum_{k_1=1}^{K_1} x_{1k_1} e_{k_1} = \left(x_{11}, x_{12}, \ldots, x_{1K_1}\right) = x_1 .$$

Similar statements can be made about components $\sum_{j=1}^{i} K_j + 1$ through

$\sum_{j=1}^{i+1} K_j$, $i = 1, \ldots, N-1$. Having shown x can be written as a convex combination of pure strategies, let us simplify the notation of (4) by writing $x = \sum_{j=1}^{M} \lambda_j e^j$ where $\{e^j , j = 1, \ldots, M\}$ are those pure strategies $e_{k_1 k_2 \ldots k_N}$ with nonzero coefficients in (4) and $\{\lambda_j , j = 1, \ldots, M\}$ are the corresponding coefficients. We want to show that all the e^j are optimal pure strategies. Suppose that e^1 is not optimal. Then there exists a state i for which

$$(5) \qquad r_i(e^1) + \beta P_i(e^1)V(x) < r_i(x) + \beta P_i(x)V(x)$$

because if not, $\geq$ would hold in (5) for all i which would imply, by the same argument used in (a), that e^1 was also optimal. Since x is optimal, we must also have

$$(6) \qquad r_i(e^j) + \beta P_i(e^j)V(x) \leq r_i(x) + \beta P_i(x)V(x) \qquad j = 2, \ldots, M .$$

Now multiply (5) by λ_1 and the equations of (6) by λ_j and sum to get $r_i(x) + \beta P_i(x)V(x) < r_i(x) + \beta P_i(x)V(x)$, a contradiction. Hence, each e^j is optimal.

Theorem:

An equilibrium point in stationary strategies exists for the discounted case of a nonzero-sum stochastic game.

Proof:

Lemmas 1 and 2 imply that the nonempty $\theta_1 \times \theta_2$ has a closed graph and is convex. Kakutani's fixed point theorem can now be applied to infer that $\theta_1 \times \theta_2$ has a fixed point and hence that the game has an equilibrium point in stationary strategies.

CHAPTER 3

AVERAGE RATE OF RETURN CASE

3.1 Introduction

The development of the average rate of return case basically follows that of the discounted case. The existence of an equilibrium point in stationary strategies is established with the aid of a linear programming formulation of the problem and assumptions on the type of Markov chain that can underly the motion of the players. the latter consideration leads to the study of three cases corresponding to the nature of the underlying chains: irreducible chains, chains with a single ergodic subchain, and chains with multiple ergodic subchains.

Once again, the analysis depends crucially on the fact that when one player uses some given strategy, the other is faced with a sequential decison process. The properties of the set of strategies that optimally answer a fixed strategy of an opposing player are used to justify the use of a fixed point theorem in order to establish the existence of an equilibrium point.

3.2 Multiple Chain Case

The objective of maximizing average rate of return in a sequential decision process is the player's desire to find a strategy x so as to maximize the column vector $G(x)$ whose ith component, $G_i(x)$, is the average reward per period when the initial state of the system is i and strategy x is employed every period. If G_i^* is defined to be the average reward per period, over all future time, when the initial state of the system is i and optimal decisions are made at the beginning of every period, then it is true that there exists a strategy that will achieve G_i^*

uniformly for all i. This can be accomplished by showing that for any randomized strategy, x, there is a pure strategy that can achieve an average reward per period vector at least as great as $G(x)$ and by using the policy-improvement routine of Howard [7]. Hence, the player's objective is a realizable one.

Recalling that when player II uses some fixed strategy $\bar{y}$ player I is faced with a sequential decision process, he will wish to maximize $G(x,\bar{y})$, while player II will seek the maximum over $y \in Y$ of $H(\bar{x},y)$, his average rate of return vector when player I uses some fixed $\bar{x}$. If X and Y are the players' strategy spaces and

$$\phi_1(\bar{y}) = \left\{ \bar{x} \;\middle|\; \max_{x \in X} G(x,\bar{y}) = G(\bar{x},\bar{y}) \right\}$$

and

$$\phi_2(\bar{x}) = \left\{ \bar{y} \;\middle|\; \max_{y \in Y} H(\bar{x},y) = H(\bar{x},\bar{y}) \right\}$$

we will again have the

Definition:

The pair of strategies (x^o,y^o) is said to be an equilibrium point if $x^o \in \phi_1(y^o)$ and $y^o \in \phi_2(x^o)$.

An example due to Gillette [4] demonstrates that in the average rate of return case, stochastic games may fail to have an equilibrium point in stationary strategies. Gillette's three state example is:

STATE	BIMATRIX GAME AND TRANSITION PROBABILITIES
1	$\begin{array}{\|c\|c\|}\hline 1,-1 \\ (1,0,0) & 0,0 \\ (0,1,0) \\ \hline 0,0 \\ (1,0,0) & 1,-1 \\ (0,0,1) \\ \hline \end{array}$
2	$\begin{array}{\|c\|}\hline 0,0 \\ (0,1,0) \\ \hline \end{array}$
3	$\begin{array}{\|c\|}\hline 1,-1 \\ (0,0,1) \\ \hline \end{array}$

$$a_{k\ell}^{i}, b_{k\ell}^{i}$$

where in state i, the entires $\left(p_{i1}^{k\ell}, p_{i2}^{k\ell}, p_{i3}^{k\ell}\right)$ are the result of players' I and II choosing actions k and ℓ respectively and represent their immediate rewards and the probability vector governing their transition out of state i.

Notice that once the players are in state 2, they remain there for all time, and each receives an average rate of return of zero. Similarly, once the players are in state 3, they are sure to remain there forever with average rates of return 1 and -1. Hence, the players are only concerned with their strategy choices in state 1, to be chosen with the intent of maximizing $G_1(x,y)$ and $H_1(x,y)$.

To show that no equilibrium point exists, note that

$$\phi_1(\bar{y}) = \begin{cases} \{((0,1;1;1))\} & \text{if } \bar{y}_{12} > 0 \\ \{((1,0;1;1))\} & \text{if } \bar{y}_{12} = 0 \end{cases}$$

$$\phi_2(\bar{x}) = \begin{cases} (1,0;1;1) & \text{if } \bar{x}_1 = (0,1) \\ \{y \mid y_{12} > 0\} & \text{if } \bar{x}_1 = (1,0) \ . \end{cases}$$

Hence, there is no pair of strategies that are mutually optimal responses, i.e., there is no pair (x^o,y^o) for which $x^o \in \phi_1(y^o)$ and $y^o \in \phi_2(x^o)$.

The crucial point to note here is the lack of continuity in the optimal responses of player I to a sequence of strategies of player II. That is to say, as $\bar{y}_{12} \to 0$, $x_1 = (0,1)$ is player I's optimal response as long as $\bar{y}_{12} > 0$. But for $\bar{y}_{12} = 0$, player I's optimal response is $x_1 = (1,0)$, certainly not the limit of a sequence of $(0,1)$'s . This condition, which arises because of the multiple chain nature of the underlying Markov chains, is what is preventing the existence of an equilibrium point. In the next section, we'll see that if the underlying chains have only a single irreducible subclass of states, the continuity described above will obtain, and an equilibrium point in stationary strategies will exist.

3.3 Irreducible Chains

In light of the multiple chain example of the last section, we see that for a general proof of the existence of an equilibrium point in the average rate of return case, we must at least restrict ourselves to games where no matter what pair of strategies, (x,y) , is chosen, the Markov chain determined, $P(x,y)$, has a single irreducible subclass of states. In the current section, an even more restrictive assumption on the chains will be made, while in the

next section, we will return to the minimal restriction mentioned above.

Assumption A:

All pairs of pure strategies, $\left(e_{k_1 k_2 \ldots k_N}, e_{\ell_1 \ell_2 \ldots \ell_N}\right)$, determine an irreducible Markov chain.

Note that this assumption is sufficient to guarantee that for any strategy pair, (x,y), $P(x,y)$ is irreducible. For the rest of this section, Assumption A is assumed to hold.

Since the chains, $P(x,y)$, are always irreducible, the starting position of the players is irrelevant with respect to average rate of return, and $G_i(x,y) = G_j(x,y)$ and $H_i(x,y) = H_j(x,y)$ for all i, j will hold for any strategy pair (x,y). So letting the scalars $g(x,y)$ and $h(x,y)$ be the players' average rates of return, we may write $g(x,y) = \pi(x,y) \cdot a(x,y)$ and $h(x,y) = \pi(x,y) \cdot b(x,y)$ where $\pi(x,y)$ uniquely satisfies $\pi(x,y) = \pi(x,y)P(x,y)$, $\sum_{i=1}^{N} \pi_i(x,y) = 1$, $\pi_i(x,y) \geq 0$.

These inner products have the interpretation of weighted averages of period rewards, i.e., in the long run, when strategy pair (x,y) is used, a proportion $\pi_i(x,y)$ of the transitions are made through state i and each time such a transition occurs, expected rewards $a_i(x,y)$ and $b_i(x,y)$ are earned. We can now simplify ϕ_1 and ϕ_2 as

$$\phi_1(\bar{y}) = \left\{ \bar{x} \mid \max_{x \in X} g(x,\bar{y}) = g(\bar{x},\bar{y}) \right\}$$

$$\phi_2(\bar{x}) = \left\{ \bar{y} \mid \max_{y \in Y} h(\bar{x},y) = h(\bar{x},\bar{y}) \right\}$$

and attempt to show that ϕ_1 and ϕ_2 have closed graphs and are convex. Once again, since a player opposing an opponent's fixed strategy is facing

the usual sequential decision process, we can concentrate our attention on the set of optimal strategies for a sequential decision process, i.e., those $\bar{x}$ which solve the nonlinear programming problem

$$
\begin{aligned}
&\text{Maximize} && g(x) = \pi(x)r(x) \\
&\text{Subject to} && \sum_{k=1}^{K_i} x_{ik} = 1 && i = 1, \ldots, N \\
&&& \pi(x) = \tau(x)P(x) \\
&&& \sum_{i=1}^{N} \pi_i(x) \geq 0 \\
&&& x \,,\ \pi(x) \geq 0
\end{aligned}
\tag{1}
$$

Manne [12] showed that (1) is equivalent to a linear program which in turn was showr by Wolfe and Dantzig [16] to be equivalent to the generalized linear program

$$
\begin{aligned}
&\text{Maximize} && rz \\
&\text{Subject to} && \sum_{i=1}^{N} Q_i z_i = \begin{pmatrix} 0_N \\ 1 \end{pmatrix} \\
&&& z \geq 0
\end{aligned}
\tag{2}
$$

where 0_N is an N-vector of zeroes and, for $i = 1, \ldots, N$, Q_i is a column in the convex polyhedron C_i generated by the K_i extreme points $Q_{ik} = \left(p_{i1}^k, \ldots, p_{ii}^k - 1, \ldots, p_{iN}^k, 1\right)$, $k = 1, \ldots, K_i$ and if

$$Q_i = \sum_{k=1}^{K_i} \lambda_{ik} Q_{ik}$$

with

$$\sum_{k=1}^{K_i} \lambda_{ik} = 1 \ , \ \lambda_{ik} \geq 0 \ ,$$

then

$$r_i = \sum_{k=1}^{K_i} \lambda_{ik} r_{ik} \ .$$

Problem (2) can be arrived at as a direct consequence of seeking randomized optimal strategies. The key is to recognize that, when in say, state i, choosing alternative k with probability $x_{ik}, k = 1, \ldots, K_i$, is, in terms of expectations, the very same thing as engaging in the single alternative represented by the appropriate probability mixture of K_i "pure" actions. This is simply a restatement of the second paragraph of Section 2.3. Hence, if P_i^k is the probability vector governing transitions out of state i if alternative k is chosen there, and r_{ik} is the associated immediate expected reward, then employing the randomized strategy x corresponds to choosing a single probability vector

$$P_i(x) = \sum_{k=1}^{K_i} P_i^k x_{ik} \ ,$$

in the convex polyhedron P_i generated by the P_i^k's, to govern transitions out of state i and to earning the immediate expected reward

$$r_i(x) = \sum_{k=1}^{K_i} r_{ik} x_{ik} \ .$$

Thus, suppressing the x's, if $P_i \in P_i, i = 1, \ldots, N$ determine the Markov chain $P = (P_1, \ldots, P_N)$ that governs state transitions, the associated

average rate of return can be found by solving $\pi = \pi P$, $\sum_{i=1}^{N} \pi_i = 1$, $\pi_i \geq 0$ to get the stationary vector π associated with P and then computing r . Since solving $\begin{pmatrix} P^T - I \\ 1 \ldots 1 \end{pmatrix} \pi = \begin{pmatrix} 0_N \\ 1 \end{pmatrix}$ for π (uniquely) leads to the computation of average rate of return, we would like to pick those members of P_j that result in the maximum average rate of return.

Now we can show how (2) arises if we let

$$C_i = \left\{ Q_i \mid Q_i = \begin{pmatrix} P_i - e_i \\ 1 \end{pmatrix} \text{ for some } P_i \in P_i \right\} .$$

Problem (2) says we would like to select Q_i from $C_i, i = 1, \ldots, N$, so as to determine a nonnegative z from $Qz = \begin{pmatrix} 0_N \\ 1 \end{pmatrix}$ that will maximize rz . But the K_i extreme points of C_i are clearly

$$Q_{ik} = \begin{pmatrix} P_i^k - e_i \\ 1 \end{pmatrix} ,$$

and the weights on the extreme points of C_i that determine

$$Q_i(x) = \sum_{k=1}^{K_i} Q_{ik} x_{ik}$$

are precisely the same weights on the extreme points of P_i needed to determine the probability transition vector resulting from strategy x , i.e., column selection from C_i is equivalent to specifying weights on the extreme points of C_i which is the same as specifying weights on the extreme points of P_i , an operation identical to choosing a randomized strategy. This establishes a $1 - 1$ correspondence between solutions to (2) and our original problem: if a randomized strategy, x , results in an average rate

of return $g(x)$, then the columns $Q_i(x) = \sum_{k=1}^{K_i} Q_{ik}x_{ik}$ and z solving $Q(x)z = \begin{pmatrix} 0_N \\ 1 \end{pmatrix}$ will result in a value of the objective $rz = g(x)$ in (2). Similarly, a set of columns Q_i, chosen as a feasible solution to (2), can be expressed as convex combinations of the extreme points of the C_i. If we let the weight on Q_{ik} necessary to express Q_i be the probability with which alternative k is chosen in state i, a strategy will result with average rate of return equal to the value of the objective in (2) for this corresponding feasible solution.

For any strategy x, the rank of $Q(x)$ is less than $N + 1$ since the first N rows sum to zero. In fact, the rank of

$$Q(x) = \begin{pmatrix} P^T(x) - I \\ 1 \ \ldots \ 1 \end{pmatrix}$$

is N as can be seen by considering the $N \times N$ matrix $\bar{Q}(x)$ obtained by arbitrarily deleting one of the first N rows of $Q(x)$. Assuming $\bar{Q}(x)$ is not of rank N implies there exists $\bar{\rho} \neq 0$ such that $\bar{Q}\bar{\rho} = 0$ (suppressing the x's). Because P is irreducible, there is a unique nonzero bounded solution $\bar{\pi}$ to

$$Q\sigma = \begin{pmatrix} 0_N \\ 1 \end{pmatrix}, \tag{3}$$

more familiarly written $\sigma = \sigma P$, $\sum_{i=1}^{N} \sigma_i = 1$ (Chung [1]). A contradiction now occurs since $\bar{\sigma} = \bar{\pi} + \bar{\rho}$ will also be a nonzero bounded solution to (3) since the row deleted from Q can be written as minus the sum of the first $N - 1$ rows of $\bar{Q}$ and the inner product of each of the first $N - 1$ rows of $\bar{Q}$ with both $\bar{\pi}$ and $\bar{\rho}$ is zero. The last row of (3) is satisfied by $\bar{\sigma}$ since the elements of $\bar{\pi}$ sum to one and those of $\bar{\rho}$ to

zero. This latter fact also assures us that $\bar{\sigma}$ cannot be zero.

We are now prepared to prove the existence of an equilibrium point for the average rate of return case with irreducible chains. Following Section 2.2 again, we will need to show the convexity and closed graph properties of ϕ_1 and ϕ_2.

Lemma 1:

$\phi_1 : Y \to 2^X$ and $\phi_2 : X \to 2^Y$ have closed graphs.

Proof:

A sufficient condition for ϕ_1 to have a closed graph is the continuity of $g(x,y) = \pi(x,y)r(x,y)$ where $\pi(x,y)$ is the stationary vector of the Markov chain $P(x,y)$, and $r(x,y)$ is the vector of immediate expected rewards when strategy pair (x,y) is employed. Since the elements of $r(x,y)$ are just bilinear terms in the x_{ik} and $y_{i\ell}$, we only have to demonstrate the continuity of $\pi(x,y)$, where $\pi(x,y)$ solves

$$\begin{pmatrix} P^T(x,y) - I \\ 1 \ \ldots\ 1 \end{pmatrix} \pi = \begin{pmatrix} 0_N \\ 1 \end{pmatrix},$$

or recalling the notation of Problem (2), $Q(x,y)\pi = \begin{pmatrix} 0_N \\ 1 \end{pmatrix}$. As a consequence of the result on the rank of $Q(x,y)$, we can arbitrarily delete one of the first N rows of $Q(x,y)$ to form $\bar{Q}(x,y)$ and write

$$\pi = \bar{Q}^{-1}(x,y)\begin{pmatrix} 0_{N-1} \\ 1 \end{pmatrix}, \tag{4}$$

always assured of the existence of $\bar{Q}^{-1}(x,y)$ for all (x,y). The elements of $\bar{Q}^{-1}(x,y)$ are just ratios of polynomials involving the x_{ik} and $y_{i\ell}$ so that the (unique) solution to (4) is, in fact, a continuous function of (x,y). An identical argument on $h(x,y)$ yields the closed graph nature of ϕ_2.

Lemma 2:

$\phi_1(\bar{y})$ can be characterized as a closed convex polyhedron whose extreme points constitute the set of pure strategies that are optimal responses to $\bar{y}$.

Proof:

As we have remarked several times in the past, when player II fixes his strategy at $\bar{y}$, player I is faced with a sequential decision process that we can put in the form of Problem (2):

$$\begin{aligned} &\text{Maximize} && a(\bar{y})z \\ (5) \qquad &\text{Subject to} && Q(\bar{y})z = \begin{pmatrix} 0_N \\ 1 \end{pmatrix} \\ & && z \geq 0 . \end{aligned}$$

Since we will want to deal with a linear program with full rank, from now on, it will be assumed that the Nth row in (5) is deleted, so that

$$P_i^k(\bar{y}) = \left(\sum_{\ell=1}^{L_i} p_{i1}^k \bar{y}_{i\ell}, \ \ldots, \ \sum_{\ell=1}^{L_i} p_{ii}^k \bar{y}_{i\ell} - 1, \ \ldots, \ \sum_{\ell=1}^{L_i} p_{i,N-1}^k \bar{y}_{i\ell} \right) .$$

$P_i(\bar{y})$ will be determined by the extreme points $P_i^k(\bar{y})$ and it, along with $Q_i(\bar{y})$ and $C_i(\bar{y})$ will have the same interpretation as before except for their reduced dimensionality. Now for any particular selection of columns

$$Q_i(x,\bar{y}) = \sum_{k=1}^{K_i} Q_{ik}(\bar{y}) x_{ik}$$

from $C_i(\bar{y}), i = 1, \ldots, N$, the system of Equations in (5) will be square and $a_i(\bar{y}) = \sum_{k=1}^{K_i} \left(\sum_{\ell=1}^{L_i} a_{k\ell}^i y_{i\ell} \right) x_{ik}$. This and earlier remarks and the irreducibility

assumption imply a 1 - 1 correspondence between strategies x and feasible nondegenerate bases $Q(x,\bar{y})$.

As in Lemma 2 of Section 2.4, the proof will be accomplished if it is shown that: (a) any convex combination of optimal pure strategies is an optimal randomized strategy and (b) if some randomized strategy is optimal, then it is a convex combination of optimal pure strategies.

(a) Suppose x^1 and x^2 are pure strategies that are optimal responses to $\bar{y}$. We want to show that $x_\lambda = \lambda x^1 + (1 - \lambda)x^2$, $0 \leq \lambda \leq 1$, is also an optimal response to $\bar{y}$. The analysis used for the development of Problem (2), and consequently Problem (5), enables us to say that x^1 and x^2 correspond to the optimal bases $Q(x^1,\bar{y})$ and $Q(x^2,\bar{y})$ and x_λ to the basis $Q(x_\lambda,\bar{y}) = \lambda Q(x^1,\bar{y}) + (1 - \lambda)Q(x^2,\bar{y})$. If μ_i is the N-vector of (optimal) simplex multipliers associated with $Q(x^i,\bar{y})$, $i = 1$, 2 , we must have $\mu_1 = \mu_2$. This is so because under nondegeneracy (which is implied here by irreducibility), the optimal multipliers μ_1 must price out the columns in $Q(x^2,\bar{y})$ zero. (This need not be true if the optimal strategy x^1 results in some transient states, for this results in degeneracy. See Section 3.4.) Hence, $a(x^2,\bar{y}) - \mu_1 Q(x^2,\bar{y}) = 0$ which implies $\mu_1 = a(x^2,\bar{y})Q^{-1}(x^2,\bar{y}) = \mu_2$. It now follows that μ_λ , the simplex multipliers associated with $Q(x_\lambda,\bar{y})$ must also equal μ_1 since $\mu_1 Q(x_\lambda,\bar{y}) = \mu_1[\lambda Q(x^1,\bar{y}) + (1 - \lambda)Q(x^2,\bar{y})] = \lambda a(x^1,\bar{y}) + (1 - \lambda)a(x^2,\bar{y}) = a(x_\lambda,\bar{y})$, so that $Q(x_\lambda,\bar{y})$ must also be an optimal basis and, therefore, x_λ an optimal strategy.

(b) Suppose x is an optimal randomized strategy with associated basis $Q(x,\bar{y})$ and optimal simplex multipliers μ . Recall that x can be written as a convex combination of pure strategies:

$$x = \sum_{k_1=1}^{K_1} \cdots \sum_{k_N=1}^{K_N} x_{1k_1} \cdots x_{Nk_N} e_{k_1 \cdots k_N} .$$

Therefore, $Q(x,\bar{y}) = \sum_{k_1=1}^{K_1} \cdots \sum_{k_N=1}^{K_N} x_{1k_1} \cdots x_{Nk_N} Q\left(e_{k_1 \cdots k_N}, \bar{y}\right)$ and

$$a(x,\bar{y}) = \sum_{k_1=1}^{K_1} \cdots \sum_{k_N=1}^{K_N} x_{1k_1} \cdots x_{1k_N} a\left(e_{k_1 \cdots k_N}, \bar{y}\right) .$$

By definition μ satisfies $a(x,\bar{y}) - \mu Q(x,\bar{y}) = 0_N$, i.e.,

$$\sum_{k_1=1}^{K_1} \cdots \sum_{k_N=1}^{K_N} x_{1k_1} \cdots x_{Nk_N} \left[a\left(e_{k_1 \cdots k_N}, \bar{y}\right) - \mu Q\left(e_{k_1 \cdots k_N}, \bar{y}\right)\right] = 0_N .$$

Since we want to show that the coefficients $x_{1k_1} \cdots x_{Nk_N}$ in this sum are positive only if $e_{k_1 \cdots k_N}$ is optimal, we break the sum into two parts, one corresponding to the set, $\mathcal{O}$, of actions $k_1, \ldots, k_N$ that are optimal, and another corresponding to the set $\bar{\mathcal{O}}$ of actions that result in nonoptimal bases. Now we have

$$\sum_{k_1,\ldots,k_N \in \mathcal{O}} x_{1k_1} \cdots x_{Nk_N} \left[a\left(e_{k_1 \cdots k_N}, \bar{y}\right) - \mu Q\left(e_{k_1 \cdots k_N}, \bar{y}\right)\right] +$$

(6)

$$\sum_{k_1,\ldots,k_N \in \bar{\mathcal{O}}} x_{1k_1} \cdots x_{Nk_N} \left[a\left(e_{k_1 \cdots k_N}, \bar{y}\right) - \mu Q\left(e_{k_1 \cdots k_N}, \bar{y}\right)\right] = 0_N ,$$

where, by the optimality of μ, $a\left(e_{k_1 \cdots k_N}, \bar{y}\right) - \mu Q\left(e_{k_1 \cdots k_N}, \bar{y}\right) \leq 0_N$ for all $k_1, \ldots, k_N$.

But under nondegeneracy, a vector of optimal simplex multipliers prices out all the columns of another basis zero if and only if the other basis is optimal. Hence, the first sum in (6) vanishes (by the "if") and at least one of the elements is negative in every vector $a\left(e_{k_1 \cdots k_N}, \bar{y}\right) - \mu Q\left(e_{k_1 \cdots k_N}, \bar{y}\right)$ in the second sum of (6) (by the "only if"). Therefore, we must have $x_{1k_1} \cdots x_{Nk_N} = 0$ for $k_1, \ldots, k_N \in \bar{\mathcal{O}}$ in order to maintain the equality

in (6), and the lemma is proven.

Theorem:

Under Assumption A, an equilibrium point in stationary strategies exists for the average rate of return case of a nonzero sum stochastic game.

Proof:

Identical to discounted case with ϕ_i replacing θ_i , $i = 1 , 2$.

3.4 Chains with a Single Ergodic Subchain

Having dealt with the two extreme cases of finite Markov chains in the last two sections, we are now left with the "in-between" case of a Markov chain that allows for some transient states, but only one irreducible subset of states. Such a chain may be taken to look like $\begin{pmatrix} A_1 & A_2 \\ 0 & P \end{pmatrix}$ where every row of A_2 has at least one positive element and the subchain P is irreducible. (The previous section assumed A_1 vacuous.) We will make an assumption analogous to that of the last section and then show that equilibrium points still exist on this middle ground, although the proofs of Section 3.3 must be modified.

Assumption B:

All pairs of pure strategies, $\left(e_{k_1 \ldots k_N}, e_{\ell_1 \ldots \ell_N}\right)$, determine a Markov chain with a single ergodic subchain.

This assumption is sufficient to guarantee that for any strategy pair (x,y) , $P(x,y)$ has a single ergodic subchain. For the rest of this section, Assumption B is assumed to hold.

Once again, the initial state of the players will have no bearing on their average rates of return which can still be expressed as

$g(x,y) = \pi(x,y)\cdot a(x,y)$ and $h(x,y) = \pi(x,y)b(x,y)$ since the solution to $\pi(x,y) = \pi(x,y)P(x,y)$, $\sum_{i=1}^{N} \pi_i(x,y) = 1$, $\pi_i(x,y) \geq 0$ remains unique. The generalized linear programming approach (Problem (5) of Section 3.3) can also be used again. The proof that the rank of $Q(x,\bar{y})$ is N given in the last section also holds under Assumption B and sustains the validity of Lemma 1 and the $1 - 1$ correspondence between strategies and feasible bases. However, the fact that a basis $Q(x,\bar{y})$ may now be degenerate (corresponding to transiency in the chain $P(x,\bar{y})$) leads to the breakdown of the convexity of $\phi_1(\bar{y})$ proved in Lemma 2. Specifically, μ may be a vector of simplex multipliers associated with an optimal basis, but may fail to price out all the columns of another optimal basis zero, since under degeneracy, dual feasibility of μ is sufficient but not necessary for optimality. Examples of the nonconvexity of $\phi_1(\bar{y})$ in the presence of transient states are easily constructed.

At this point, the natural thing to do is to turn to a generalization of Kakutani's fixed point theorem that would weaken the convexity requirement on $\phi_1(\bar{y})$, since, as remarked above, the closed graph property of $\phi_1(\bar{y})$ still obtains under Assumption B. Such a generalization has been given by Debreu [2] in an adaptation of a fixed point theorem of Eilenberg and Montgomery [3]. Here, the convexity requirement is replaced by the requirement that $\phi_1(\bar{y})$ be contractible, the topological equivalent of convexity. Schweitzer [14] has shown that $\phi_1(\bar{y})$ may be represented as a convex set with convex protuberances, but such a set may in general fail to be contractible, and it is not clear how to use the properties of the special structure at hand to show that $\phi_1(\bar{y})$ is, in fact, contractible. However, Schweitzer's decomposition of $\phi_1(\bar{y})$ leads to another consideration which results in a way that circumvents the current difficulty.

Suppose we denote by $\psi_1(\bar{y})$ that subset of $\phi_1(\bar{y})$ that is the convex "core" of $\phi_1(\bar{y})$, i.e., $\phi_1(\bar{y}) - \psi_1(\bar{y})$ composes the protuberances of $\psi_1(\bar{y})$. Convexity will no longer be a problem if we can show that $\psi_1(\bar{y})$ has a closed graph, for then we can still use Kakutani's theorem to prove the existence of a fixed point for $\psi_1(\bar{y})$. This will be good enough to insure the existence of an equilibrium point because elements of $\psi_1(\bar{y})$ (and $\psi_2(\bar{x})$) are still optimal. Interestingly enough, in Schweitzer's decomposition of $\phi_1(\bar{y})$, the convex set from which convex protuberances emenate is the set of all optimal strategies $\bar{x}$ with associated basis $Q(\bar{x},\bar{y})$ that determine simplex multipliers $\mu(\bar{x},\bar{y}) = a(\bar{x},\bar{y})Q^{-1}(\bar{x},\bar{y})$ which are dual feasible, i.e., price out all the extreme points (and hence, all columns) of $C_i(\bar{y})$ nonpositively for all i.

More formally, define:

$$\psi_1(\bar{y}) = \Big\{ \bar{x} \;\Big|\; \max_{x \in X} g(x,\bar{y}) = g(\bar{x},\bar{y}) \;,\; a_{ik}(\bar{y}) - \mu(\bar{x},\bar{y})Q_{ik}(\bar{y}) \leq 0 \;\; \forall \;\; i \;,\; k$$

$$\text{where } \mu(\bar{x},\bar{y}) = a(\bar{x},\bar{y})Q^{-1}(\bar{x},\bar{y}) \Big\} \;.$$

$\psi_2(\bar{x})$ is analogously defined for the generalized linear program that arises when Player II has to find an optimal response to Player I's use of $\bar{x}$.

Lemma 1:

$\psi_1 : Y \to 2^X$ and $\psi_2 : X \to 2^Y$ have closed graphs.

Proof:

Let $\{y^n\}$ be a sequence of Player II's strategies converging to y^o and $\{x^n\}$ the sequence of corresponding optimal responses of player I, i.e., $x^n \in \psi_1(y^n) \forall n$, with $x^n \to x^o$. We have to show that x^o is an optimal response to y^o. Since $x^n \in \psi_1(y^n) \;\forall\; n$, $a_{ik}(y^n) - \mu(x^n,y^n)Q_{ik}(y^n) \leq 0 \;\forall$ i, k, n. The continuity of $a_{ik}(y)$, $\mu(x,y)$ and $Q_{ik}(y)$ is assured as

in Lemma 1 of the last section with Assumption B crucial here in guaranteeing the existence of $\mu(x^n,y^n) = a(x^n,y^n)Q^{-1}(x^n,y^n)$ for all n. Hence, $a_{ik}(y^o) - \mu(x^o,y^o)Q_{ik}(y^o) \leq 0$ giving $x^o \in \psi_1(y^o)$. $\psi_2(\bar{x})$ has a closed graph by the same argument.

Lemma 2:

$\psi_1(\bar{y})$ and $\psi_2(\bar{x})$ are convex.

Proof:

Theorem 10, Schweitzer [14].

Theorem:

Under Assumption B, an equlibrium point in stationary strategies exists for the average rate of return case of a nonzero sum sotchastic game.

Proof:

Identical to discounted case with ψ_i replacing θ_i, $i = 1, 2$.

3.5 Extensions

The above theorem is easily generalized in two directions. The same development follows if we have an n-person stochastic game where an n-tuple of strategies, $s_1^o, \ldots, s_n^o$, one for each player, is an equilibrium point if for any player, say the ith, s_i^o maximizes player i's average rate of return when opposing the fixed strategies s_j^o, $i \neq j$, of the other $n - 1$ players. We can appropriately define the correspondences $\psi_1, \ldots, \psi_n$, each being a subset of the optimal solutions to a sequential decision process. Consequently, Lemmas 1 and 2 hold for all ψ_i, so that $\psi = \psi_1 \times \ldots \times \psi_n$ has a fixed point and an equilibrium point exists.

Another generalization concerns the underlying law of motion governing the players' state transitions. If we assume that the players' joint choice of actions in a state not only determines their immediate expected rewards and transition probabilities, but also specifies a probability distribution of the time to the next transition (that may depend on the state to which a transition is made), then a semi-Markov process underlies the motion of the players (rather than a Markov chain when the above mentioned probability distributions are degenerate at a unit time) and their objectives become maximization of long run average rate of return per unit time. Howard [8] showed that for the one player case of this set-up, an optimal policy for the sequential decision process only depends on the probability distribution of transition times through their first moments.

In addition, the problem can be formulated as a generalized linear program just as in the Markov chain case. All that is needed is to modify (5) in Section 3.3 by changing the extreme points of $C_i(\bar{y})$ to

$$Q'_{ik}(\bar{y}) = Q_{ik}(\bar{y}) \frac{1}{\tau_{ik}(\bar{y})}$$

where

$$\tau_{ik}(\bar{y}) = \sum_{\ell=1}^{L_i} \tau^i_{k\ell} \bar{y}^{\ell}_i$$

and $\tau^i_{k\ell}$ is the mean time spent in state i if the players use pure strategy pair (e_k, e_ℓ). Lemmas 1 and 2 remain unchanged for the appropriately modified ψ_1 and ψ_2 so that the existence theorem also holds for this more general case.

3.6 An Equivalence Theorem

Just as it is possible to pose a sequential decision process as a generalized linear programming problem, it is possible to cast a two-person nonzero-sum stochastic game under Assumption A as a programming problem, although a nonlinear one in this case. Let

$E(y) = \left(Q_{11}(y), \ldots, Q_{1K_1}(y), \ldots, Q_{N1}(y), \ldots, Q_{NK_N}(y)\right)$ be the matrix of extreme points of the convex polyhedra $C_i(y)$, $i = 1, \ldots, N$, determined by player II's strategy y and from which player I is to choose columns in problem (5) of Section 3.3. $F(x)$ is analogously defined to be the matrix of extreme columns of the generalized linear program faced by player II when player I uses strategy x. Let

$e(y) = \left(a_{11}(y), \ldots, a_{1K_1}(y), \ldots, a_{N1}(y), \ldots, a_{NK_N}(y)\right)$ and

$f(x) = \left(b_{11}(x), \ldots, b_{1L_1}(x), \ldots, b_{N1}(x), \ldots, b_{NL_N}(x)\right)$ be the vectors of associated rewards.

It is easily seen that the linear program

$$\text{(1)} \qquad \begin{aligned} &\text{Maximize} && e(y)w \\ &\text{subject to} && E(y)w = \begin{pmatrix} 0_{N-1} \\ 1 \end{pmatrix} \\ & && w \geq 0 \end{aligned}$$

is equivalent to Section 3.3's problem (5). Given a solution w to (1) above, we can get a solution to (5) by letting the weight on column $Q_{ik}(y)$ needed to express $Q_i(y)$ be

$$x_{ik} = \frac{w_{ik}}{\sum_{j=1}^{K_i} w_{ij}}$$

and letting

$$z_i = \sum_{k=1}^{K_i} w_{ik} .$$

Similarly, given a solution to (5), letting $w_{ik} = z_i x_{ik}$, where x_{ik} is again the weight on $Q_{ik}(y)$ needed to express $Q_i(y)$, solves (1) above for a given solution to (5). Note that

(2) $\qquad x_{ik} > 0$ if and only if $w_{ik} > 0$.

Against player I's x, player II faces the linear program

$$\begin{aligned} &\text{Maximize} && f(x)z \\ (3) \qquad &\text{subject to} && F(x)z = \begin{pmatrix} 0_{N-1} \\ 1 \end{pmatrix} \\ & && z \geq 0 \end{aligned}$$

Again, we have

(4) $\qquad y_{i\ell} > 0$ if and only if $z_{ik} > 0$.

Lemma:

Under Assumption A, a necessary and sufficient condition for the strategy pair (x^o, y^o) to be an equilibrium point is that there exist μ^o and ν^o such that

(i) $\quad [e(y^o) - \mu^o E(y^o)]x^o = 0$

(ii) $\quad [f(x^o) - \nu^o F(x^o)]y^o = 0$

(iii) $\quad e(y^o) - \mu^o E(y^o) \leq 0$

(iv) $\quad f(x^o) - \nu^o F(x^o) \leq 0$

(v) $\quad x \in X$

(vi) $\quad y \in Y$

Note:

μ^o and ν^o must be the vectors of simplex multipliers associated with x^o and y^o and problems (1) and (3) above.

Proof:

The existence of a μ^o satisfying (i) and (iii) is guaranteed by (2) and the necessary and sufficient conditions for optimality of the non-degenerate (Assumption A) linear programming problem (1). Symmetrically, the existence of a ν^o satisfying (ii) and (iv) is guaranteed by (4) and the optimality conditions of the linear programming problem (3).

Theorem:

Under Assumption A, a necessary and sufficient condition for the strategy pair (x^o,y^o) to be an equilibrium point is that x^o, y^o, and some μ^o, ν^o solve the nonlinear programming problem

$$
\begin{aligned}
&\text{Maximize} && \{[e(y) - \mu E(y)]x + [f(x) - \nu F(x)]y\}^{\dagger} \\
&\text{subject to} && e(y) - \mu E(y) \leq 0 \\
(*) \qquad && && f(x) - \nu F(x) \leq 0 \\
&&& x \in X \\
&&& y \in Y .
\end{aligned}
$$

---

†Letting $\delta_{ij} = 0$ for $i \neq j$ and $\delta_{ij} = 1$ for $i = j$, the objective can be written

$$\sum_{i=1}^{N} \sum_{k=1}^{K_i} \sum_{j=1}^{N-1} \sum_{\ell=1}^{L_i} \left(a_{k\ell}^{i} - p_{ij}^{k\ell} + \delta_{ij}\right)\mu_j x_{ik} y_{i\ell} +$$

$$\sum_{i=1}^{N} \sum_{k=1}^{K_i} \sum_{j=1}^{N-1} \sum_{\ell=1}^{L_i} \left(b_{k\ell}^{i} - p_{ij}^{k\ell} + \delta_{ij}\right)\nu_j x_{ik} y_{i\ell} - (\mu_N + \nu_N) .$$

Proof:

Sufficiency: Let x^o , y^o , μ^o , ν^o solve (*) . We will show that (i) through (vi) in the lemma hold. Feasibility guarantees (iii) through (vi) and

$$(5) \qquad [e(y^o) - \mu^o E(y^o)]x^o + [f(x^o) - \nu^o F(x^o)]y^o \leq 0 .$$

But by the existence theorem of Section 3.3, there exist $\bar{x}$, $\bar{y}$, $\bar{\mu}$, $\bar{\nu}$ satisfying (i) through (vi). Hence, there exists a feasible solution to (*) with the value of the objective equal to zero so that equality must hold in (5). Finally, equality in (5) and (iii) through (vi) imply (i) and (ii) are satisfied. Now we can apply the sufficiency part of the lemma to infer that (x^o,y^o) is an equilibrium point.

Necessity: Let (x^o,y^o) be an equilibrium point with associated simplex multipliers μ^o and ν^o . Then (i) through (vi) hold implying x^o , y^o , μ^o , ν^o are feasible for (*) and, in fact, solve (*) since zero is achieved for an objective that is nonpositive for all feasible solutions to (*) .

Several comments can be made about the equivalence theorem. For a one state problem ($N = 1$) , the theorem reduces to a theorem given by Mangasarian [11] for bimatrix games. Another note of interest is the fact that the average rate of return for players I and II associated with equilibrium point (x^o,y^o) is μ_N^o and ν_N^o respectively, a consequence of the duality theorem of linear programming since $\mu^o\begin{pmatrix}0_{N-1}\\1\end{pmatrix} = \mu_N^o$ and $\nu^o\begin{pmatrix}0_{N-1}\\1\end{pmatrix} = \nu_N^o$, $\begin{pmatrix}0_{N-1}\\1\end{pmatrix}$ being the right-hand side of both players' linear programs. Finally, only the sufficiency part of the theorem holds under Assumption B since, under degeneracy, some equilibrium points (x,y) may have associated simplex multipliers that are not dual feasible, i.e., fail to

satisfy (iii) and/or (iv).

The complexity of the equivalent nonlinear program indicates that it may be most difficult to find its solutions. An intuitively appealing approach is the following iterative scheme: at the nth iteration, x and y are fixed at some $x(n)$, $y(n)$. The associated optimal simplex multipliers, $\mu(n)$ and $\nu(n)$, to the linear programs determined by $x(n)$ and $y(n)$ are then found. Then (*) is solved with μ and ν fixed at $\mu(n)$ and $\nu(n)$ respectively. This determines $x(n+1)$ and $y(n+1)$ to be used for the $(n+1)$st iteration. Hopefully, $x(n)$ and $y(n)$ converge to an equilibrium pair. But there is no guarantee that this process will converge to a solution to (*) (and hence an equilibrium point) since the possibility exists that the scheme will get hung-up around a set of variables $\hat{x}$, $\hat{y}$, $\hat{\mu}$, $\hat{\nu}$ where $\hat{\mu}$ and $\hat{\nu}$ are simplex multipliers determined by $\hat{x}$ and $\hat{y}$ and $\hat{x}$ and $\hat{y}$ solve (*) for the fixed $\hat{\mu}$ and $\hat{\nu}$. It is interesting to note, however, that to find the (unique) equilibrium point of a zero-sum stochastic game, this procedure works and is precisely the same algorithm as the one given by Hoffman and Karp [6] for determining the value and optimal strategies for zero-sum games.

3.7 Possibilities for Further Research

Both applied and theoretical problems related to the results given here present possibilities for further research. Many authors have discussed the formulation of an infinite horizon periodic review inventory model as a sequential decision process. The state at the beginning of a period is the inventory on hand and the actions available to the system operator correspond to the level up to which he orders while the expected immediate rewards correspond to the expected net revenue for the period: expected sales minus expected ordering, holding, and shortage costs. The probability distribution of demand and a particular choice of order levels determine the transition

probabilities that govern state transitions.

Now consider two operators of inventory systems who stock the same item. If a demand is unsatisfied by the first operator, it is reasonable to assume that this demand may revert to the second operator rather than be backordered with the first, and thus affecting the demand pattern and reward structure of the second. A similar statement is true about a demand unsatisfied by the second operator. Hence, the policies of the two operators may be considered as a nonzero-sum stochastic game since the reward structure and transition probabilities clearly depend on the operators' joint actions. Rational operators (in the game theoretic sense) of such inventory systems will tend to seek equilibrium operating strategies.

Consideration of such a problem leads directly to two possible extensions of a theoretical nature. A characterization of the set of all equilibrium points (perhaps making use of the equivalence theorem) of a nonzero-sum stochastic game would be helpful in resolving situations where one equilibrium point is preferred by one operator and a second equilibrium point by the other, or a situation where one equilibrium point is better (for both players) than all others. A second extension would deal with the problem of partial state information, i.e., a player has some idea about the state he's in (for example, *his* inventory level) but lacks total state information (for example, his *opponent's* inventory level).

Other areas for further work readily follow from the consideration of extensions to basic game theory and sequential decision problems, e.g., co-operative games, various solution concepts, and allowing for a countable number of states and actions.

REFERENCES

[1] Chung, K. L., MARKOV CHAINS, Springer-Verlag, Inc., New York, (1967).

[2] Debreu, G., "A Social Equilibrium Existence Theorem," Proceedings of the National Academy of Sciences, U.S.A., Vol. 38, pp. 886-893, (1952).

[3] Eilenberg, S. and D. Montgomery, "Fixed Point Theorems for Multi-Valued Transformations," American Journal of Mathematics, Vol. 68, pp. 214-222, (1946).

[4] Gillette, D., "Stochastic Games with Zero Stop Probabilities," CONTRIBUTIONS TO THE THEORY OF GAMES, Vol. III, pp. 179-187, Princeton, (1957).

[5] Hadley, G., NONLINEAR AND DYNAMIC PROGRAMMING, Addison-Wesley Publishing Company, Inc., (1964).

[6] Hoffman, A. J. and R. M. Karp, "On Nonterminating Stochastic Games," Management Science, Vol. 12, pp. 359-370, (1966).

[7] Howard, R., DYNAMIC PROGRAMMING AND MARKOV PROCESSES, M.I.T. Press, Cambridge, Massachusetts, (1960).

[8] Howard, R., "Research in Semi-Markovian Decision Structures," Journal of the Operations Research Society of Japan, Vol. 6, pp. 163-199, (1964).

[9] Kakutani, S., "A Generalization of Brouwer's Fixed Point Theorem," Duke Mathematical Journal, Vol. 8, pp. 457-458, (1941).

[10] Luce, R. D. and H. Raiffa, GAMES AND DECISIONS: INTRODUCTION AND CRITICAL SURVEY, John Wiley and Sons, (1957).

[11] Mangasarian, O. L., "Equilibrium Points of Bimatrix Games," Journal of SIAM, Vol. 12, pp. 778-780, (1964).

[12] Manne, A., "Linear Programming and Sequential Decisions," Management Science, Vol. 6, pp. 259-267, (1960).

[13] Nash, J., "Equilibrium Points in N-Person Games," Proceedings of the National Academy of Sciences, U.S.A., Vol. 36, pp. 48-49, (1950).

[14] Schweitzer, P., "Randomized Gain-Optimal Policies for Undiscounted Markov Renewal Programming," Institute for Defense Analyses Report, Arlington, Virginia.

[15] Shapley, L. S., "Stochastic Games," Proceedings of the National Academy of Sciences, U.S.A., Vol. 39, pp. 1095-1100, (1953).

[16] Wolfe, P. and G. B. Dantzig, "Linear Programming in a Markov Chain," Operations Research, Vol. 10, pp. 702-710, (1962).

Unclassified

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)

University of California, Berkeley

2a. REPORT SECURITY CLASSIFICATION

Unclassified

2b. GROUP

3. REPORT TITLE

NONZERO-SUM STOCHASTIC GAMES

4. DESCRIPTIVE NOTES (Type of report and, inclusive dates)

Research Report

5. AUTHOR(S) (First name, middle initial, last name)

Philip D. Rogers

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
April 1969	38	16

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
Nonr-222(83) b. PROJECT NO NR 047 033 c. Research Project No.: RR 003 07 01 d.	ORC 69-8 9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)

10. DISTRIBUTION STATEMENT

This document has been approved for public release and sale; its distribution is unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
NONE	MATHEMATICAL SCIENCE DIVISION

13. ABSTRACT

SEE ABSTRACT.

DD FORM 1 NOV 65 1473 (PAGE 1)

S/N 0101-807-6811

Unclassified

Security Classification

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Nonzero-Sum Games						
Sequential Decision Process						
Equilibrium Points						
Stochastic Game						